



**HAL**  
open science

## STAGE - Data Management Plan n°1

Jeanne Fras, Clarisse Bardiot

► **To cite this version:**

Jeanne Fras, Clarisse Bardiot. STAGE - Data Management Plan n°1. Université Rennes 2. 2025.  
hal-04879289

**HAL Id: hal-04879289**

**<https://hal.science/hal-04879289v1>**

Submitted on 10 Jan 2025

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# From Stage to Data

STAGE project - Data Management Plan (DMP)

First version

2024/06/24

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement n° 101097091 - STAGE).



## Plan Details 4

Revision Board	4
Specifications	4
Associated Documents	6
Plan Purpose/Scope	6
General Overview	7
Data Summary	7
Data Organization	8
Ethics Monitoring	11
Risk Assessment	11

## Project Details 13

Abstract	13
Context	15
Research Outputs	15
DMP Contributors	16

## Project Data 17

Saving, Publishing, Archiving	17
Estimated Size	17
FAIR Data	18

## STAGE - Research Products Description 20

Summary	20
Performing Arts Ontology	20
Data Set: Digitization And Study Of Avignon's Festival Programs	20
Data Set: 20 Years+ Photographies By Christophe Raynaud De Lage	22
Data Set: Study Of 15 Performances' Creative Processes At Avignon's Festival	23
Collection Of The 15 Creative Processes	23
Data Visualizations	25
Arvest	25
FAIR Data and Resources	26
Performing Arts Ontology	26

Data Set: Digitization And Study Of Avignon's Festival Programs	27
Data Set: 20 Years+ Photographies By Christophe Raynaud De Lage	28
Data Set: Study Of 15 Performances' Creative Processes At Avignon's Festival	29
Collection Of The 15 Creative Processes	30
Data Visualizations	31
Arvest	32

## Appendixes 33

Appendix 1: Data Protection Impact Assessment (DPIA)	33
--	----

# Plan Details

Plan title: DMP of the project "STAGE: From Stage to Data, the Digital Turn of Contemporary Performing Arts Historiography »

## Revision Board

Name	Date	Modifications
Jeanne FRAS	2024/04/15	Creation
Clarisse Bardiot	2024/05/03	Revisions
Jeanne FRAS	2024/05/07	Corrections
Alessia Smaniotto	2024/05/13	Revisions
Jeanne Fras	2024/06/24	Corrections

## Specifications

Fields of science and technology (from OECD classification):

Art (arts, history of arts, performing arts, music), Computer and information sciences

Language: eng

Creation date: 2024-04-05

Last modification date: 2024-06-24

License Name: Creative Commons Attribution 4.0 International

URL: <http://spdx.org/licenses/CC-BY-4.0.json>



## Associated Documents



- Website : <https://stage-to-data.huma-num.fr/>
- HAL-SHS Collection : <https://hal.science/STAGE>
- Hypothèses Blog : <https://paa.hypotheses.org>

## Plan Purpose/Scope

The aim of this document is to outline the strategy and protocols established for the collection, usage, and protection of data throughout the STAGE project. The Data Management Plan (DMP) explains the management decisions made to ensure these aspects, as well as the preservation of the research products once the project ends. The protocols described were developed following the Findable, Accessible, Interoperable, and Reusable (FAIR) principles, striving to make the data and deliverables "as open as possible, as closed as necessary." Finally, this DMP aims to present how the procedures essential for compliance with the ethical requirements related to the guidelines of the European Research Council will be respected.

This DMP will evolve during the project and will be updated at three separate time frames, describing the changes, challenges, and solutions to problems encountered. During the first six months of the STAGE project, the general guidelines will be defined and set, along with the identification of the different deliverables and data sets expected from the project; this is the first version of this document.

The STAGE DMP includes information and details on:



- Data collection: what data, data origin, contributors, data processing, generated data
- Data organization: repositories, naming conventions, data set relations, metadata description
- Data accessibility: formats, publication, licenses
- Data preservation: archiving, backups, data protection
- Ethical monitoring within the project

This DMP was written using the DMP OPIDoR tool and the ERC DMP templates available online.

## General Overview

In this first section, we will present the general protocols established, which apply to all datasets and research products, before going into detail in a second section.

---

## Data Summary

The STAGE project will explore and cross several existing corpora of different origins, leading to the creation of datasets of both homogeneous and heterogeneous composition. At the same time, research activities are planned to create corpora as well (creative process studies), which are labeled as collections in this document.

In general, four categories can be distinguished:

- Existing collections (already public and published data)
- Created collections (collected and organized data)
- Management data (insights on data management in STAGE, DMP)
- Results data (research outputs, deliverables)

Types of deliverables expected as outputs of the project are:

- Datasets (e.g., Avignon's Festival Programs, Christophe Raynaud de Lage's pictures, Creative Processes reorganized)
- Interactive resources (e.g., Data Visualizations)
- Models (e.g., Performing arts ontology)
- Software (e.g., Arvest)
- Collections (e.g., Creative processes, companies' hard drives and archives)
- Texts (e.g., Data papers)



- Workflows (e.g., DMP, Management Data)

---

## Data Organization

Rules concerning the internal organization of documents and files manipulated during data collection and research have been established. To ensure good management – meaning to avoid duplicates, confusion in versioning documents, or loss of a file – a naming convention and a clear directory hierarchy (Figure 1) have been set. Each repository level is supported by a README text file indicating the expected content, so that every user of the storage space uploads files to the correct directory.

The naming convention aims to keep a clear track of document versions, their chronology, and to quickly identify the content and type of a file using its name. Our naming convention was built based on examples from French universities displayed online and the advice of the Maison des Sciences de l'Homme en Bretagne's (MSHB) seminars.

Composed of:

- The date of creation/modification (YYYY-MM-DD)
- The project it's related to (2-4 letter acronym)
- The type of document (slides, template, etc.) (2-3 letter acronym)
- A short name
- The version of the document (Vx, VSR, VBAT)

Following these requirements, an example of document naming shows as follow:

- A slideshow created on April 4, 2024, to present the Christophe Raynaud de Lage dataset, modified three times, will be named:  
2024-04-04\_CRDL\_SLD\_Present-DS\_V3.odt
- A template, received on March 12, 2024, for an authorization of use of image, written in French, will be named:  
2024-03-12\_ADM\_TEMP\_Right-Image-Consent-FR.docx



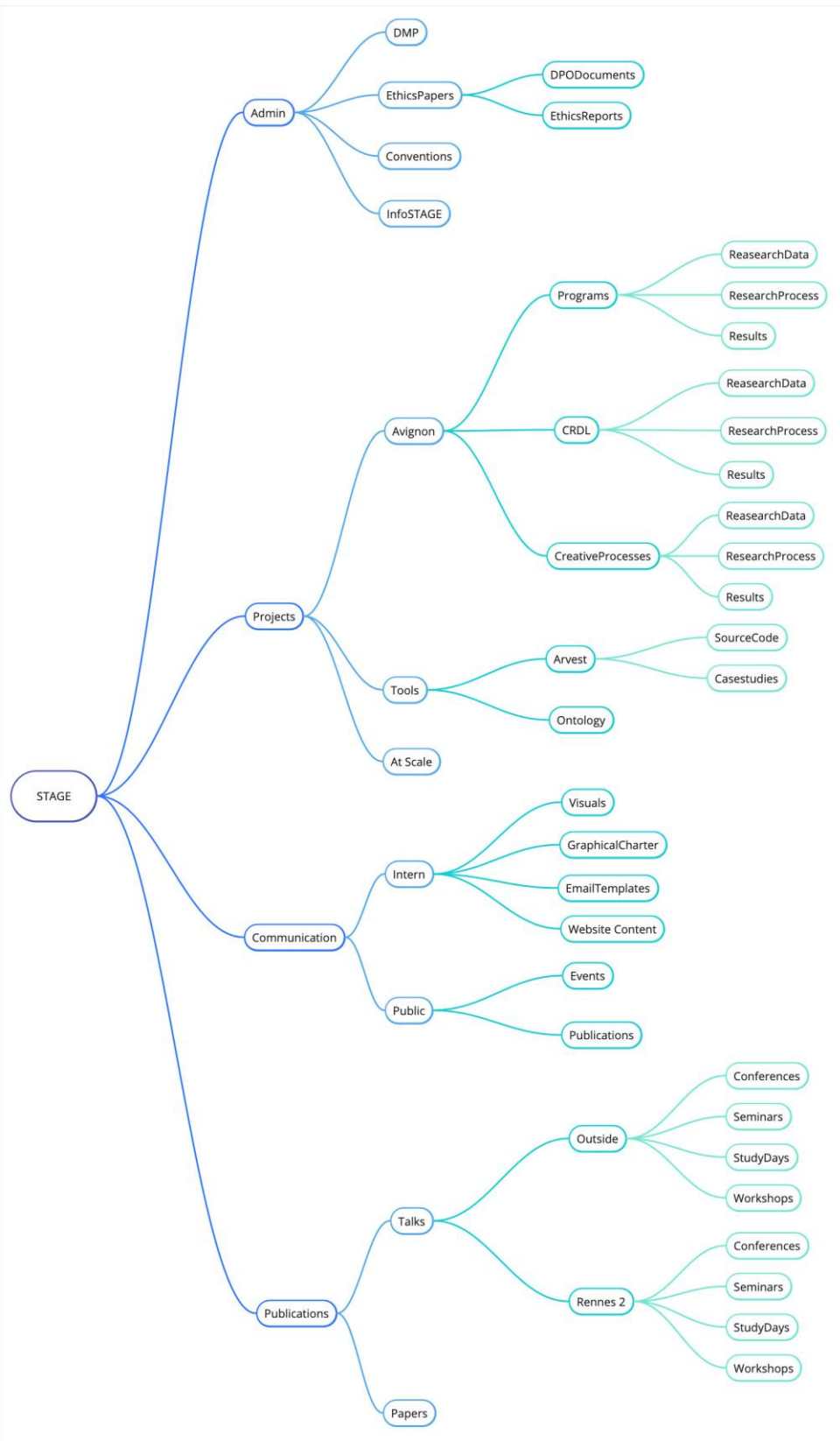


Figure 1 : File hierarchy  
STAGE - 2024

---

## Ethics Monitoring

The STAGE project research objectives leads to the collection of personal data, which implies enhanced monitoring of the legal aspects of collecting, studying, and sharing data during and after the project. Supplementing this DMP, which will represent the summary of guidelines established and followed, the Principal Investigator (PI) and Data Engineer are working in close collaboration with the Data Protection Officer (DPO) at the host institution of STAGE (Université Rennes 2) to comply with current legislation such as the General Data Protection Regulation (GDPR). They are also creating documents for external contributors of the project (creative processes analysis). Guidelines published by the Commission Nationale de l'Informatique et des Libertés (CNIL) are also key resources used to establish the project's boundaries and freedoms.

Moreover, ethics reports are to be sent every year to the European Research Council, which will be verified by the STAGE project's Ethics Advisor. The project submitted a request for the certification of the Comité d'Éthique de la Recherche of it's host institution (Université Rennes 2).

---

## Risk Assessment

STAGE acknowledges that using AI-related tools to process personal data raises ethical and bias-related questions that will be addressed with the participation of the DPO during the establishment of the relevant analysis. The project also acknowledges that the consent of its participants is essential for the right to collect, process, publish, and archive of the research products (the creative processes collection and dataset). Therefore, the project will be diligent in explaining these processes to the participants. This DMP is linked to a Data Protection Impact Assessment (DPIA). According to the [Ethics and Data Protection \(2021\)](#) document, those situations present a higher ethic risk, hence the need of a DPIA (Appendix 1, p.29).

To avoid any conflict or dissatisfaction from our contributors, the process of interviews to present the project, its objectives, and obstacles, the consent forms, as well as publishing and archiving terms of service, will be co-designed and established together. This will allow any participant to be fully aware of and personally engaged with all agreements made. Additionally,

the collection of personal data from participants will rely on each person sending the data they consent to share. Participants will be encouraged to provide any data that addresses the issues and questions STAGE aims to explore. Prior explanations of these issues will be accompanied by examples of documents that could help the project.

Regarding biases, AI analysis is planned to be separated into three steps, corresponding to different degrees of supervision in the algorithm's learning process. Starting with an unsupervised scenario, there will be a second step of fine-tuning the results and algorithms used, followed by supervised learning. During this process, we will pay specific attention to data, model and interpretation bias.

# Project Details

Project title: STAGE – From Stage to Data, the Digital Turn of Contemporary Performing Arts Historiography

Acronym: STAGE

## Abstract

Digitized or born-digital, digital sources are one of the most significant challenges facing performing arts historiography: they are fragile (technological obsolescence), multimodal (texts, images, videos, programs), and numerous (big data). Think, for example, about the 20,000 photographs taken each year by the official photographer of the Avignon Festival; or the 15,000 files documenting the creation process of one work. Digital traces are the future of performing arts studies and our new primary sources. The shift to digital not only transforms the nature of our sources but also reshapes how research is conducted and its outcomes. How can we interpret digital traces? How can we “make them talk”? What new insights can we draw from them? STAGE tackles these questions head-on by offering a theoretical and methodological framework situated at the confluence of history, epistemology, and digital humanities. Drawing from the Avignon festival collection and employing interdisciplinary methodologies, we aim to renew performing arts studies and demonstrate the profound significance of digital traces in both preserving and analyzing our cultural heritage. By leveraging the advancements in digital humanities and artificial intelligence, we uncover new insights into the historical and aesthetic dimensions of *mise en scène* since WWII. Through the two prisms of resurgence and collaboration, STAGE endeavors to reveal creation contexts and networks, aesthetic reminiscences and creative process models. This approach lays the groundwork for what we term “performing arts analytics. »

STAGE, spanning from January 2024 to December 2028, delves into a pivotal moment of transformation: the transition of traditional sources into digital traces. This paradigm shift alters the nature of historical sources, operating on both hermeneutic and epistemological levels.

Digital technology transforms traces into data, inviting us to reconsider traditional research approaches in performing arts studies. To build these methods and demonstrate their potential, we will rely first on a corpus that has not yet been explored at this scale and which offers a particularly fertile field to approach such questions: the archives of the Avignon Festival, before opening to broader corpora in a second phase in order to scale up our results and expand our analysis. Bridging qualitative and quantitative research approaches, STAGE focuses on three intermediate objectives:

1) To visualize performing arts, by creating a network using data from programs to showcase interactions among thousands of individuals, different forms of artistic and technical collaborations as well as the context of creation over time. This approach, drawing on actor-network theory, will lead to an updated perspective of the rich and complex context of contemporary European stagings. However, a crucial challenge lies in the absence of a universal standard for describing performing arts. STAGE plans to address this by creating an ontology to enhance description and facilitate interoperability with other collections and data sets.

2) To reveal staging intertextuality. Photographs and videos are crucial traces in performing arts, enabling "distant viewing". By applying intertextuality to stage imagery, we unveil complex relationships with past performances, identifying resurgences and connections. Advances in computer vision facilitate the development of an iconology of performing arts, tracing aesthetic networks. STAGE will contribute to the training and development of computer-vision algorithms dedicated to performing arts.

3) To model creative processes, taking into account not only the rehearsals but also all the data produced by all the team members. The computational analysis of performances' creative process, particularly their collaborative dimension, presents significant challenges in performing arts studies. Digital traces, often ignored in actual research on creative processes, offer new opportunities by capturing the entire creation process, from initial ideas to premiere. By developing a multimodal environment to collect and analyze data from 15 performances,

STAGE will renew the study of creative processes. This approach will unveil unique insights into the diversity of performing arts practices.

## Context

Funding: European Commission: <https://doi.org/10.3030/101097091>

Start date: 2024-01-01

End date: 2028-12-31

Partners: Université Rennes-II <https://ror.org/01m84wm78>

## Research Outputs

- Performing Arts Ontology (Model)
- Data-Set : Digitization and study of Avignon's festival programs (data set)
- Data-Set : 20 years+ Photographies Christophe Raynaud de Lage (data set)
- Data-Set : Study of 15 creative processes for representations at Avignon's festival (data set)
- Collection Of The 15 Creative Processes (Collection)
- Data Visualizations (Interactive resource)
- Arvest (Software)



# DMP Contributors

Roles	Name	ORCID	Affiliation
PI	BARDIOT Clarisse	0000-0002-8126-8249	Université Rennes 2
Host	Université Rennes 2	0000 0001 2152 2279	Université Rennes 2
DMP Manager	FRAS Jeanne	0009-0003-2633-8253	Université Rennes 2
Ethics Advisor	SMANIOTTO Alessia	0000-0001-9018-3776	EHESS

Contact for data (Arvest, Ontology, Creative Processes, DS Programs Avignon, DS Creative process, DS CRDL, Data Vizs) : FRAS Jeanne ([jeanne.fras@univ-rennes2.fr](mailto:jeanne.fras@univ-rennes2.fr)).

# Project Data

## Saving, Publishing, Archiving

Depending on the formats, subjects, and content of the data and research products, we have identified appropriate platforms to save, archive, and publish our data. All platforms tend to follow the open science initiative and are mostly managed by public services provided to researchers.

The chosen services are:

- Storage: To handle 'hot data', we use safe sharing file spaces with simultaneous editing, access level management, and versioning management features (RESANA [French government] and SHAREDOCS [Huma-Num]).
- Publication: DOI attributions are a priority in the publication process. Therefore, the choice was made depending on an automatic DOI attribution feature, as well as a metadata description service (NAKALA [Huma-Num] and GitHub + Zenodo [CERN and Microsoft]).
- Archive: The platforms chosen to publish our data are linked to archive services, where the 'cold data' will be stored (CINES [French government] and Software Heritage [INRIA]).

## Estimated Size

In the current state, the total size of the data collected, processed, and generated can only be estimated. Existing corpora that will be explored and tools partially developed are unlikely to evolve significantly, representing the most reliable indicator of the minimum total size.

For the collections to be created, based on previous similar studies, we can approximately calculate the size.



- BnF's programs collection: 1 TB
- Christophe Raynaud de Lage's personal photograph collection: 10 TB in 2024, more to be added
- Arvest: 5 GB
- Creative processes: 15 TB, 1 TB per company studied

In total, the project expects at least 31 TB of deliverables and data.

## FAIR Data

Applying FAIR principles during STAGE is a significant challenge but a priority. The team will mostly use open-source tools, and the interoperability and reusability of the outputs is a central objective. Regarding open access publication and data sharing, as stated earlier, most services used are public services or aim to facilitate the discovery and open archiving of research products.

An important contributor in this project, which also hosts the STAGE website, is the IR\* (Infrastructure de Recherche Étoile) Huma-Num. Providing free hosting service, online safe collaborative storage space, a metadata description in Dublin Core format, and ensuring great sustainability and availability for its services, Huma-Num is essential for applying FAIR principles.

The main FAIR issue we will encounter is in the context of the creative process study. Performing arts workers and creators tend to use very specific tools to support and construct their productions. These creations are documented by schemas and scores created on applications using non-open formats. Despite this, the data are crucial for documenting the processes and cannot be ignored. Digital traces and contemporary digital history of performing arts are important questions STAGE aims to study, explaining the distinction between the Creative Processes Dataset and Creative Processes Collection. Despite the lack of interoperability of some files, this collection needs to be published and archived in its original state. We will clearly document the copyright status of each piece of data, specifying the

licensing terms and any restrictions on use. We may need to implement access controls or restrictions to comply with copyright laws. For instance, providing access to lower-resolution images for public use while restricting high-resolution versions to authorized users. We will clearly communicate any access restrictions and usage terms to users, possibly through a terms of use agreement they must accept before accessing the data.

Publications concerning the projects and events surrounding the state of the art in STAGE's context will also be published in open access, using two platforms:

- Hypothèses: News, articles, recruitment, events information
- HAL-SHS: Scientific publications

# STAGE - Research Products

## Description

### Summary

---

#### Performing Arts Ontology

Project Acronym: Ontology

Project Number: 1

#### Summary

Since no universal standard for the description of performing arts exists – several models exist but aren't fully interoperable – the need to create such a model was identified for this project. To resolve the complex task of visualizing the network of thousands of performance contributors and to provide a glimpse of the context of contemporary European stagings, STAGE will create an ontology. This ontology aims to allow existing databases to connect and share data.

---

#### Data Set: Digitization And Study Of Avignon's Festival

#### PAF

Project Acronym: DS PAF

Project Number: 2

#### Summary

As part of the study of the history and evolution of the performing arts milieu, STAGE is focusing on a corpus that transcribes most of this history since 1947: the programs of the Festival d'Avignon. Archives of these documents are accessible online in public databases, and the aim of this data set is to group and study all of them to help visualize performing arts history.



Three main collections are of interest:

- The Bibliothèque Nationale de France (BnF)
- Archives du Spectacle (<https://lesarchivesduspectacle.net>)
- Théâtre Contemporain (<https://www.theatre-contemporain.net/>)

The data in this data set are collected through agreements between the management structures of these databases, the BnF, and Artcena, settled by jointly written conventions. All formats of this data set are to be open, and it represents approximately 1 TB maximum.

---

## Data Set: 20 Years+ Photographies By Christophe Raynaud De Lage

- Project Acronym: DS CRDL
- Project Number: 3
- Summary

The Festival d'Avignon is a very popular and famous event, holding a central position in representing the current and past European performing arts status, broadcasting various mises en scène and creations every year. These have been captured by Christophe Raynaud de Lage, the official photographer of the event for the past 20 years. His collection represents a tremendous amount of images, growing by 20,000 photos per year, making it a significant point of interest in the STAGE project.

The data will be processed using AI, more specifically pose detection algorithms, to enlighten patterns or variations in performances and mise en scène. First, tests using non supervised learning will be made, to detect possible biases. Then fine tuning the algorithm will better the results, and eventually, if necessary, a supervised learning will be set up.

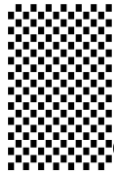
As with other collaborations, the collection and study of this corpus is supported by a convention. The transfer of this data set is made using hard drives. This data set is expected to represent around 10 TB.

---

# Data Set: Study Of 15 Performances' Creative Processes At Avignon's Festival

Project Acronym: DS CP

Project Number: 4



## Summary

To gain insights into the plurality of performing arts practices, STAGE will analyze creative processes of performers. This data set is a structured and curated group of documents, exchanges, and notes witnessing this creation by documenting all the steps it entails. It represents a reworked and organized version of the collections surrounding creative processes.

This data set, as well as the collection it is built from, involves data legally characterized as personal information. This challenge is addressed by donations from participants, who are performing artists, of their hard drives or personal computer archives. These donations, like other contributions, are supported by a convention to contextualize and frame the limits of the processing and publication, as well as consent forms and information notices. The participants are allowed to sort out files from their personal hard drives that they judge non pertinent or purely personal beforehand, with the aid and indications on what is of interest for STAGE. All of these are created following guidelines from STAGE's host institution's Data Protection Officer (DPO). The size of this data set is estimated at 15 TB.

This data is the result of the creative processes collection (described below) undergoing different entity extraction processes, in order to realize a topic modeling of the different ideas used to create a performance. All the documents will be processed by AI, NLP algorithms for texts and Computer Vision for images, to annotate researched content, bringing to light the characteristics of a creative process and its impact on the final form.

---

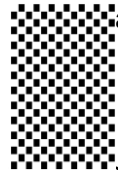
## Collection Of The 15 Creative Processes

Project Acronym: Collection CP

Project Number: 5



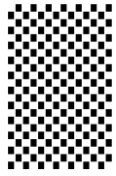
## Summary

ative Processes Collection corresponds to the raw version of the similarly named data a trace of the initial organization of files, versioning management, and naming habits, all provide insights into the study of creative processes. Like the data set, the collection of these data will mostly be by donation, with collaborating companies providing their personal computer hard drives to be explored and studied. It is also estimated to represent around 15 TB. Finally, the collection will, like the data set, be limited and supported by precise consent forms and information notices.

This dataset represents a significant part of the STAGE project objectives, which is to demonstrate how digital traces are essential elements of contemporary performing arts history and often constitute the only records of the creation of these performances. Programs and other archival objects merely mark an event, serving as witnesses to the existence of a performance. In contrast, these digital traces allow researchers to fully retrace the origins, ideas, inspirations, and overall development of a performance.

---

## Data Visualizations



Project Acronym: VIZS

Project Number: 6

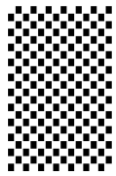
Summary

Resulting from some of the data collected and data sets created during STAGE, graphic visualizations of performing arts intertextuality and networks are composed to deepen our analysis. Through analysis using tools such as NLP or Computer Vision, these illustrations will enable distant reading of the different corpora studied. Consequently, statistics or visible patterns might materialize, raising new questions and observable trends.

Depending on the results and technical requirements, these Data Visualizations may take two forms: interactive resources, most likely available online, or images, most likely available on STAGE's website. The size of these images/interactive resources is yet to be estimated.

---

## Arvest



Project Acronym: Arvest

Project Number: 7

Summary

Arvest is a free and open-source web app that allows the annotation of videos and multimodal documents (audio, images, documents, etc.), developed in the context of STAGE. It is a Mirador extension capable of building networks of IIIF manifests by annotating manifests with other manifests. Arvest will be hosted by Huma-Num at the end of its development. In total, this app represents approximately 5 GB.

# FAIR Data and Resources

---

## Performing Arts Ontology

### 1 – Making data findable

Once developed in Protégé, a free, open-source ontology editor and framework for building intelligent systems, the ontology will be published on GitLab and assigned a DOI through Zenodo. Additionally, data papers presenting the ontology will be published on Nakala, automatically receiving a DOI. Reflection on implementation in Heurist, a free, powerful, flexible databasing and web publishing tool, is ongoing.

### 2 – Making data openly accessible

The ontology will be fully open source. Aiming to establish a norm for the description of performing arts, it must be as accessible as possible. No sensitive or personal data will be part of this deliverable, allowing it to be publicly available. Documentation and metadata description will be available alongside the ontology description itself.

Naming: ONT

### 3 – Making data interoperable

The ontology will be either be constructed using the CIDOC Conceptual Reference Model or the FRBRoo, two standards in describing relations and collections of cultural heritage. Metadata concerning the publication of this ontology will follow the Dublin Core guidelines.

### 4 – Increasing data reuse

The sharing of this model is intended to generalize habits within the community and contribute to the development of contemporary studies of digital traces in performing arts. The reusability and sustainability of this production are essential, with no embargo placed or foreseen after publication. This production will be fully open source and licensed under Creative Commons.

5 – Allocation of resources and data security

Although the model's composition involves collecting personal data, no sensitive or personal data will be published. The platforms used for publication are secured and sustainable. Nakala and Huma-Num will archive the data sets or resources published on their platform using archiving by CINES. Therefore, no funds are required for maintaining or archiving this and other research products.

---

## Data Set: Digitization And Study Of Avignon's Festival

### Programs

1 – Making data findable

This data set will be published on Nakala, assigning it a DOI in the process. Nakala provides a metadata writing feature to describe the published data set. The metadata will be published in the same place as the data set.

2 – Making data openly accessible

The programs of Festival d'Avignon data set consists of already public and published data. No access restrictions are foreseen, and the data set will be fully open and accessible. Any Python code or XML related to the collection of these data will also be published on Nakala in the same collection as the data set. This also applies to its metadata description.

Naming: PAF

3 – Making data interoperable

Metadata will be structured and described following the Dublin Core standard. Scripts used to query the databases will be written in Python or SPARQL, both open formats.

4 – Increasing data reuse

The data set aims to be fully open and freely accessible, licensed under Creative Commons. Given the heterogeneous data, massive correction tools are not very relevant. To ensure quality, tools for versioning management and history of modifications preservation will be used, such as Arkindex or Open Refine. READMEs will be available in every directory of the storage space to preserve consistency.

5 – Allocation of resources and data security

Publishing on Nakala incurs no costs for publishing or archiving, as it is directly linked to CINES, ensuring free archiving of published data sets. The corpora studied are public, but data will be secured in online storage tools like ShareDocs.

---

## Data Set: 20 Years+ Photographies By Christophe Raynaud De Lage

1 – Making data findable

Due to copyright legislations, many pictures are already public or will be published by their owner. Remote consultation of unpublished photos is under discussion, with possible access to the project's NAS for researchers meeting specific criteria.

2 – Making data openly accessible

Following the 'Open as possible, closed as necessary' principle, this data set falls under copyright laws and will not be openly accessible on Nakala. However, its metadata will be published on Nakala.

Naming: CRDL

3 – Making data interoperable

The data set's metadata will follow the Dublin Core standard. Most collected data will be transformed to the open format TIFF before reception, but some may remain in their original format: RAW.

4 – Increasing data reuse

Currently, this data-set falls under the copyright legislation, therefore most of the photographies that aren't available at the BnF won't be published by the project, implying no accessibility or reusability. Despite, one of STAGE's goal is to help achieving the complete archiving of this collection, this is still subject of conversations. Quality assurance will involve tools like Duplicate Cleaner and ReNamer.

5 – Allocation of resources and data security

Backups will rely on hard drives and the use of a NAS. The conservation, security, and

archiving will likely be managed by Université Rennes 2, with the NAS stored with the rest of their digital storage. No costs are predicted for preservation beyond the NAS purchase.

---

## Data Set: Study Of 15 Performances' Creative Processes At Jonon's Festival

### 1 – Making data findable

Publications around this data set, including metadata descriptions and the data set itself, will be made on Nakala and assigned a DOI.

### 2 – Making data openly accessible

This data set aims to be fully open. The main obstacle is the use of personal data, which requires clear and informed consent from participants to comply with GDPR. Consent forms and information notices were created based on guidelines from STAGE's host institution's DPO and the ethics committee of ERC. An ethics compliance certificate from Université Rennes 2's research ethics committee (CER) was granted for collecting and publishing this data set.

Naming: CP

### 3 – Making data interoperable

Metadata will follow the Dublin Core standard. Formats in performing arts digital tools vary, and many are not open or interoperable. STAGE compiled a list of applications and formats used by artists, including Max/MSP, Millumin, Autocad, Ableton, and their respective file formats (.amp, .amxd, .dfx, .dwg). Preserving original document formats is part of the archiving protocol.

### 4 – Increasing data reuse

The data set will be fully reusable, no embargo is foreseen, as all participants will consent to publication. Differences between the data set and the collection lie in file reorganization, hierarchy, naming conventions, and duplicates. Tools like ReNamer and Duplicate Cleaner will process data from the collections. This data set is licensed under Creative Commons.

5 – Allocation of resources and data security

Data transfers will be made by hand, passing hard drives directly to the STAGE team. Secured Huma-Num's online storage solution, ShareDocs, will be used for storing personal data ranging from professional email to Curriculum Vitae.

---

## ection Of The 15 Creative Processes

1 – Making data findable

All publications surrounding the collections of creative processes will be published on Nakala with their metadata descriptions, giving the collection a DOI.

2 – Making data openly accessible

The collection will be fully open and freely accessible, with metadata descriptions published on Nakala. The main obstacle is the use of personal data, requiring clear and informed consent from participants to comply with GDPR. Consent forms and information notices were created based on guidelines from STAGE's host institution's DPO and the ethics committee of ERC. An ethics compliance certificate from Université Rennes 2's research ethics committee (CER) was granted for this process.

Naming: CP[Name\_of\_the\_company]

3 – Making data interoperable

Similar to the data set, specific formats from professional tools like Max/MSP, Millumin, or Ableton will be included, with original formats preserved as part of the protocol.

4 – Increasing data reuse

No embargo is foreseen, as all participants will consent to publication. No quality assurance protocols are required for processing this collection, as its raw aspects are part of STAGE's analysis. Duplicates, file names, and hierarchy are integral to understanding the creative process. This collection is licensed under Creative Commons.

5 – Allocation of resources and data security

Data transfers will be made by hand, passing hard drives directly to the STAGE team.

Secured Huma-Num's online storage solution, ShareDocs, will be used for storing personal data.

---

## a Visualizations

### 1 – Making data findable

Data Visualizations will be published on Nakala, with a DOI assigned and metadata described using Nakala.

### 2 – Making data openly accessible

These visualizations will be fully open and accessible online, with metadata published on Nakala. STAGE will primarily use Gephi and D3 for creating these visualizations.

Exports in open formats will be provided, and users are recommended to use compatible tools to view the original production.

Naming: VIZS

### 3 – Making data interoperable

Visualization exports will be available as PDF, CSV, or SVG, all open formats, and in specific formats like GDF or GraphML for direct manipulation in Gephi. Metadata will follow the Dublin Core standard.

### 4 – Increasing data reuse

Visualizations will be fully open, with no embargo. Open Refine and Python scripts will ensure quality and replicability, handling history of modifications and exports of manipulation scripts. These visualizations are licensed under Creative Commons.



5 – Allocation of resources and data security

Backups will be made using hard drives stored at Université Rennes 2. Long-term preservation is ensured through Nakala, responsible for transferring, archiving, and curating publications at CINES.

---

## Arvest

1 – Making data findable

The source code of Arvest will be fully accessible on GitLab, along with its metadata description. A DOI will be attributed to it through Zenodo.

2 – Making data openly accessible

Arvest aims to be completely open source. Documentation for both users and developers will be provided at the same location as the source code.

Naming: AVST

3 – Making data interoperable

This web app is built using REACT 18 (JavaScript) and Django (Python), both fully open tools based on open formats. Metadata will be described following the Dublin Core standard.

4 – Increasing data reuse

Arvest is developed under a GPL license, ensuring its reusability and adaptability by the community.

5 – Allocation of resources and data security

Publishing the code on GitLab and assigning a DOI through Zenodo also facilitates archiving the code at Software Heritage, free of cost. For ongoing maintenance and upgrades, Arvest will have a designated community forum built on Discourse.

# Appendixes

## Appendix 1: Data Protection Impact Assessment (DPIA)

Identify the need for a DPIA

The main aim of STAGE is to enhance the importance of digital traces in the study of performing arts historiography. To achieve this analysis, STAGE will notably collaborate with companies during their creative process, by exploiting any data which stems from it. This includes mails, Curriculum Vitae, photos and several other personal data.

Those data will be processed using AI, and following the principle 'as open as possible, as closed as necessary', they are collected to compose a public and published dataset.

According to the [Ethics and Data Protection \(2021\)](#) document, those situations present a higher ethic risk, hence the need of this document.

Describe the processing

Describe the nature of the processing:

STAGE will collect the data by harvesting all the hard drives of the computer owned by members of the (performing arts) company. Those will be stored at the TGIR Huma-Num, in a secured and limited access online storage space. Back-ups will be made on hard drives.

Data studied will firstly only be shared with researchers of STAGE, before being published as a research product.

Data collected and processed will be available to all the research team of STAGE.

The need of this DPIA lies in the collection, processing and publication of personal data.

Describe the scope of the processing:

No special category of data will be either processed or collected (i.e. sensitive data). The objective is to fully exploit any digital documents and traces that the creative process involves. Nevertheless, the high risk involving those data made STAGE decide the limitation in the collection of data, by allowing collaborators to submit only the documents they judge useful and pertinent. In this judgement, STAGE will lead and recommend the selection of collaborators, by giving examples of object and document of interest.

The data will not be preserved for longer than 10 years from the end of STAGE (2028 December).

The precise number of individuals participating is currently unknown, since it'll depend on which company will collaborate with the project.

The area of study is most likely limited to France, though some artists and companies could be participating from abroad (in the limit of E.U. Territory)

Describe the context of the processing:

The individuals involved in the project are considered as contributors, and as such, they will fully be participating in the definition of the Terms of Use of their data, as well as the consent forms that will be distributed. They are in control of what they will submit to the project.

All information concerning collection processing and publication will be known by all the collaborators. No sensible group are involved.

The project is aware of the current legislation surrounding the protection of personal data (GRDP) and is dedicated to respect in the limits of it.

STAGE is currently waiting on the certification of it's host institutions (Université Rennes 2) Comité de l'éthique de la recherche (ethical research monitoring committee).

Describe the purposes of the processing:

The main objective is to achieve topics modeling throughout the corpora, in order to reveal the chronological appearances of ideas and the cited and visible influences in the process of creation of a performance. This will be achieved using NLP algorithms to reveal mentions of ideas or names, and Computer Vision algorithms to find tendencies, influences and motifs.

Annotations are to be made automatically on each type of documents, and a network of documents to better visualize the creation process will be resulting of those analysis.

This process allow a medium scale fast analysis, and will occasionally bring up further questions, which STAGE intend to answer.

In general, a system of donation is under consideration, to open this research to a broader corpus.

### Consultation process

Contributing companies are considered as essential parts of STAGE's stakeholders, and are given a place and a parole in the decision making surrounding the design of the data collection, publication and archiving. Exchanges on the subject of consent forms, conventions and terms of use of their data will be made jointly, to insure the best comprehension and knowledge of our contributors throughout the research process.

This project is currently working with an ethics advisor and regularly consult it's host institution's DPO.

### Assess necessity and proportionality

The collection of data is based on the clear and informed consent of collaborators. No analysis and publication are to be made without it. Every detail of communication, development or creation is welcomed as a research resource for STAGE, as every file, duplicate or email can reveal informations on the creative process.

All the necessary informations, concerning the project, the processing and the intents, as well as the right and liberties of any participant will be made clear in a information notice distributed beforehand.

Furthermore, each contributor will have total control over what they submit to the project, while being guided through a list of potential sources of interest by the team. Except for some photos or videos, no other data concerning them will be collected or processed. STAGE is setting up a digital secured space to submit the concerned data.

The project is accompanied by a DPO and an Ethics Advisor, to ensure our compliance with legal limitations, and team members of STAGE (Jeanne FRAS, Clarisse BARDIOT) will remain available if any question or rights are supported.

Identify measures to reduce risk				
Collection of personal data				
Risk	Options to reduce or eliminate risk	Effect on risk	Residual risk	Measure approved
No consent	Consent form Give the control of the donation to the participant.	Reduced	Medium	Yes

Processing of personal data				
Risk	Options to reduce or eliminate risk	Effect on risk	Residual risk	Measure approved
No consent Biases in the process	Consent form Unsupervised, finetuning, then supervised.	Reduced	Medium	Yes

Publication of personal data				
Risk	Options to reduce or eliminate risk	Effect on risk	Residual risk	Measure approved
No consent Use of 'droit à l'effacement / l'oubli'	Consent form Set up of Terms of Use commonly with participant and archiving service.	Reduced	Low	Yes



Sign off and record outcomes		
Item	Name/date	Notes
Measures approved by:	Clarisse Bardiot 04/06/24	Completion deadline : 02/09/24
Residual risks approved by:	Clarisse Bardiot 04/06/24	/
DPO advice provided:	Inès Rauturier 16/04/24	Second meeting with PI, discussions on ethics obstacles
Summary of DPO advice: Reliance on the consent form and information notice, advice on content of these documents.		
DPO advice accepted or overruled by:	Clarisse Bardiot 11/06/24	/
Comments: /		
This DPIA will be kept under review by:	Alessia Smaniotto 13/06/24	/



Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or ERC. Neither the European Union nor the granting authority can be held responsible for them.